University of
CINCINNATI

Q&A session for Final Exam

Zhaohu(Jonathan) Fan

08/03/2021

- Review of topics covered:
  - Chapter 1
  - Chapter 2
  - Chapter 3
  - Chapter 4
  - Chapter 5
- Exam

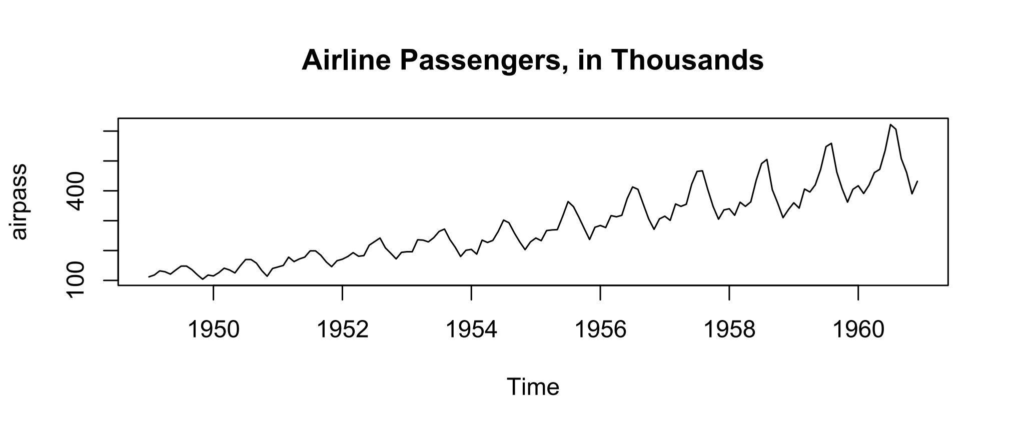
Chapter 1

Time series data

- Time series data form an ordered sequence of numbers, corresponding to an object like quantities, prices,counts, observed at or over a particular point **in time**.

- The "Airline Passengers" data set is an example of **time series data**.

```
plot(airpass)
title(main="Airline Passengers, in Thousands")
```



Time series patterns

- Describing a time series: trend, seasonality, cycles, changing variance, unusual features.
  - **Trend**: pattern exists when there is a long-term increase or decrease in the data.
  - **Seasonal** : pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week).
  - **Cyclic** : pattern exists when data exhibit rises and falls that are *not of fixed period* (duration usually of at least 2 years).

Seasonal or cyclic?

Differences between seasonal and cyclic patterns:

- seasonal pattern constant length; cyclic pattern variable length
- average length of cycle longer than length of seasonal pattern
- magnitude of cycle more variable than magnitude of seasonal pattern

The timing of peaks and troughs is predictable with seasonal data, but unpredictable in the long term with cyclic data.

Visualizing time series data in R

- Visualization is good practice to be able to understand the properties of the data.
  - Most time series coming from official data sources provide recordings at a regularly spaced set of such time points, such as every day, week, month, quarter, or year; this interval is called the **frequency** of the time series.
- A time series is stored in a `ts` object in R:
  - `ts` objects and `ts` function

For observations that are more frequent than once per year, add a `frequency` argument.

E.g., monthly data stored as a numerical vector `z`:

```r
y ← ts(z, frequency=12, start=c(2003, 1))
```

Chapter 2

Chapter 2

Correlation

- Correlation: measure of linear relationships (**a measure of the direction and strength of the relationship between two variables**)

- We use Pearson's r as a measure of the linear relationship between two quantitative variables. In a sample, we use the symbol r . In a population, we use the Greek letter ("rho"). Pearson's r can easily be computed using R.

- The correlation coefficient, rho, measures linear relationships: Ranges over [-1, +1]
    - A value of +1 indicates a perfect positive (upward sloping) linear relationship between the two variables.
    - A value of -1 indicates a perfect negative (downward sloping) linear relationship between the two variables.
    - A value of zero indicates no linear relationship between the two variables.

Interpret correlation coefficient

- Correlation coefficient is comprised between -1 and 1:
  - -0.86 indicates a strong negative correlation : this means that every time x increases, y decreases.
  - 0 means that there is no association between the two variables (x and y).
  - 0.87 indicates a strong positive correlation: this means that y increases with x .

Note: The closer r is to 0 the weaker the relationship and the closer to +1 or -1 the stronger the relationship (e.g., r=-0.98 is a stronger relationship than r=+0.78); the sign of the correlation provides direction only.

Chapter 3

Residual diagnostics

- Residuals in forecasting: difference between observed value and its fitted value: $e_t = y_t - \hat{y}_{t|t-1}$.

- Assumptions:

    - (1) $\{e_t\}$ are **uncorrelated**. If they aren't, then information left in residuals that should be used in computing forecasts.
    - (2) $\{e_t\}$ have **mean zero**. If they don't, then forecasts are biased.

- Useful properties (for prediction intervals):

    - (1) $\{e_t\}$ have **constant variance**.
    - (2) $\{e_t\}$ are **normally distributed**.

Autocorrelation (ACF)

Covariance and **correlation**: measure extent of **linear relationship** between two variables (Y and X).

Autocovariance and **autocorrelation**: measure linear relationship between **lagged values** of a time series y.

We measure the relationship between:

- y_t and y_{t-1}
- y_t and y_{t-2}
- y_t and y_{t-3}
- etc.

ACF of residuals

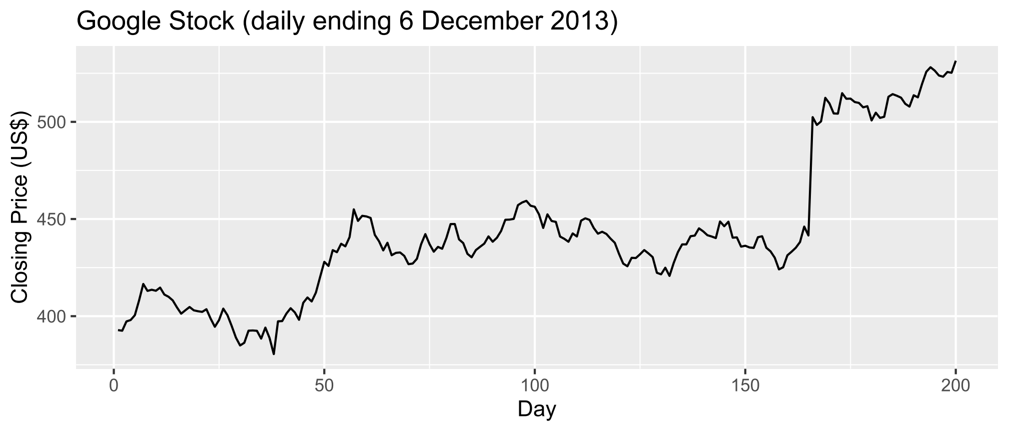
- We assume that the residuals are white noise (**uncorrelated**, mean zero, constant variance). If they aren't, then there is information left in the residuals that should be used in computing forecasts.

- So a standard residual diagnostic is to check the ACF of the residuals of a forecasting model.

- We expect these to look like white noise.

Data Example: Google stock price

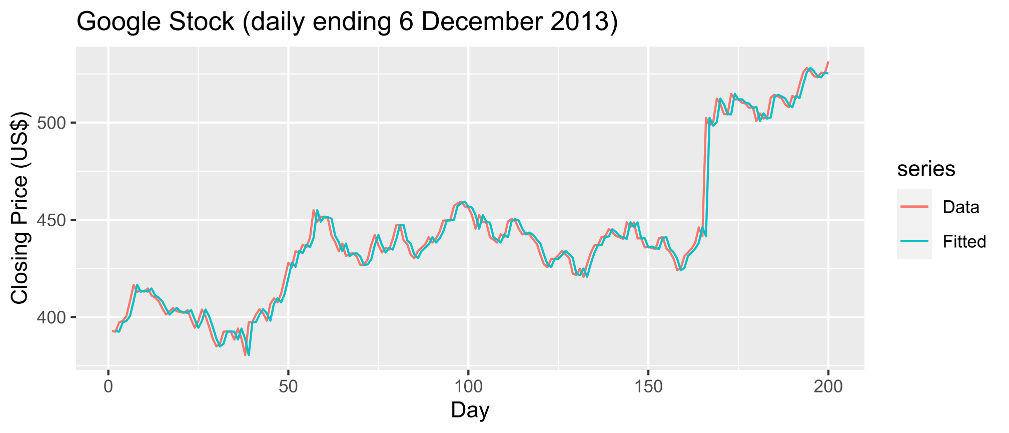
```r
autoplot(goog200) +
  xlab("Day") + ylab("Closing Price (US$)") +
  ggtitle("Google Stock (daily ending 6 December 2013)")
```



Data Example (cont'd)

- Naïve model

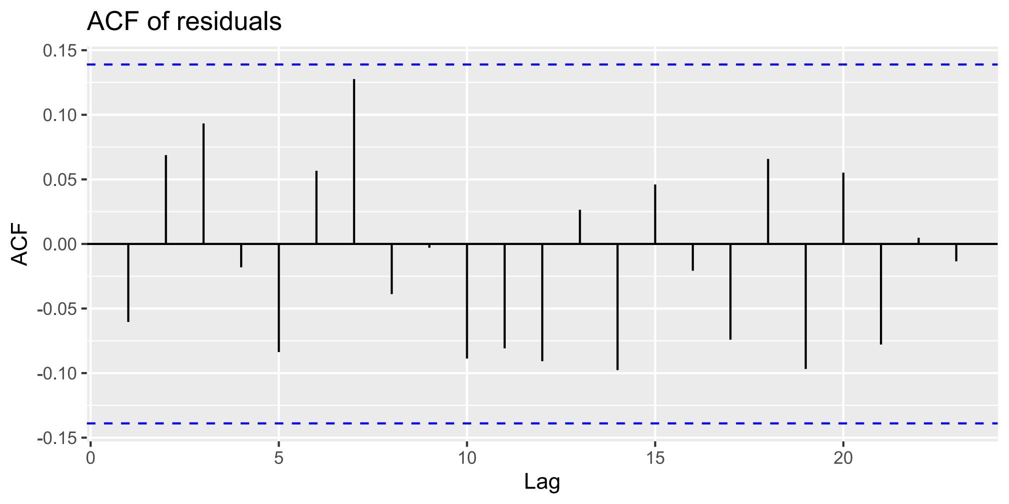
```
fits ← fitted(naive(goog200))
autoplot(goog200, series="Data") +
  autolayer(fits, series="Fitted") +
  xlab("Day") + ylab("Closing Price (US$)") +
  ggtitle("Google Stock (daily ending 6 December 2013)")
```



Data Example (cont'd)

- check if the residuals are white noise (**uncorrelated?**).
- We expect each autocorrelation to be close to zero.

```r
res ← residuals(naive(goog200))
ggAcf(res) + ggtitle("ACF of residuals")
```

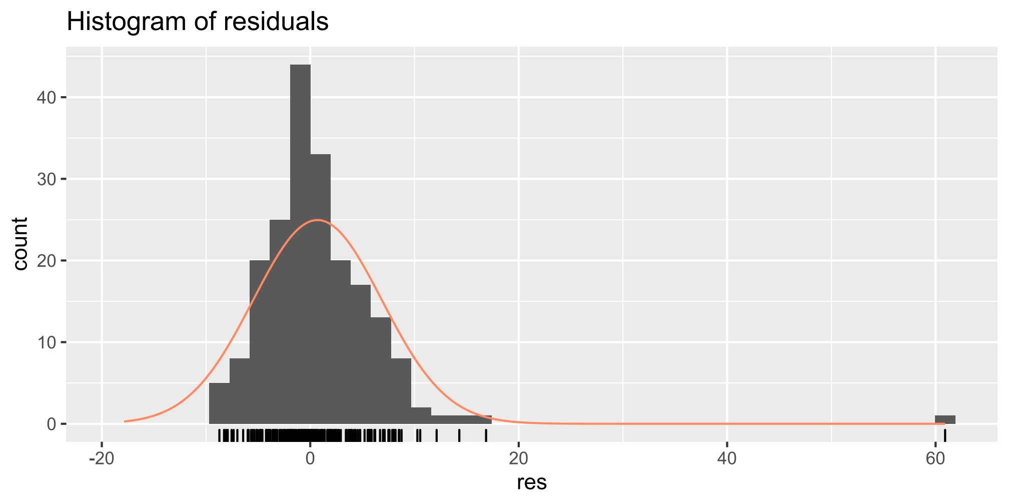


- There is no significant correlation in the residuals series.The vertical lines (lags) don't pass through horizontal blue lines. This means the autocorrelations are not (statistically) significantly different from zero.

Data Example (cont'd)

- check if the residuals are white noise (**normally distributed+ mean zero?**).

```
gghistogram(res, add.normal=TRUE) +
  ggtitle("Histogram of residuals")
```

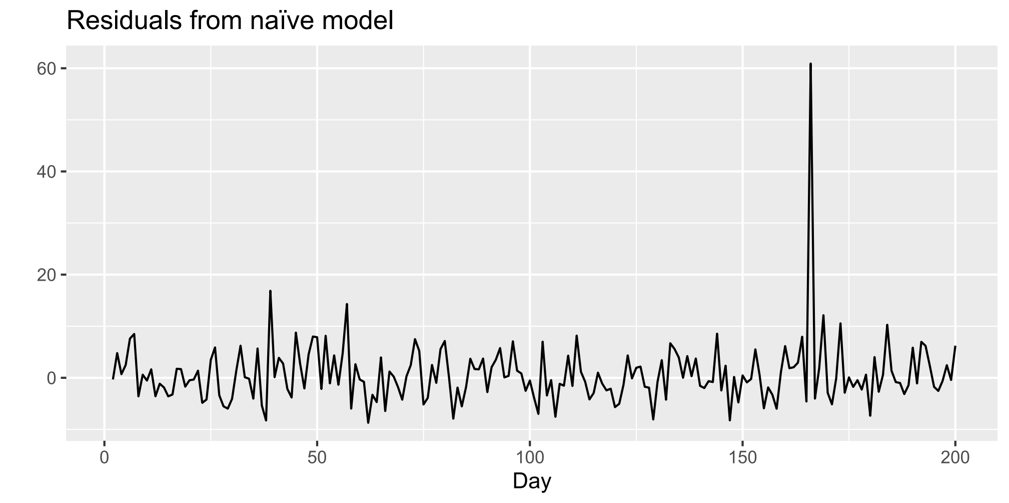


- The mean of the residuals is close to zero.
- The histogram suggests that the residuals may not be normal. The right tail seems a little too long, even when we ignore the outlier.

Data Example (cont'd)

- check if the residuals are white noise (**constant variance?**).

```
autoplot(res) + xlab("Day") + ylab("") +
  ggtitle("Residuals from naïve model")
```



- The time plot of the residuals shows that the variation of the residuals stays much the same across the historical data, apart from the one outlier, and therefore the residual variance can be treated as constant.

Data Example (cont'd): Conclusion

- Forecasts from this method (Naïve model) will probably be quite good,
- but prediction intervals that are computed assuming a normal distribution may be inaccurate.

Ljung-Box test

- In general, the Ljung-Box test (a test of autocorrelation of the residuals) is defined as:
    - H_0 (null hypothesis): the autocorrelations come from a white noise series.
    - H_a (alternative hypothesis): the autocorrelations do not come from a white noise series.

- Ljung-Box test is based on
    - $Q^* = T(T+2)\sum_{k=1}^{h}(T-k)^{-1}r_k^2$ where h is max lag being considered and T is number of observations.
    - large values of Q^* suggest that the autocorrelations do not come from a white noise series.

- By convention: $h = 10$ for non-seasonal data, $h = 2m$ for seasonal data, where m is the period of seasonality.

- Better performance, especially in small samples.

Ljung-Box test (cont'd)

- If data are white noise, Q^* has χ^2 distribution with $(h - K)$ degrees of freedom where $K =$ no. parameters in model.
- For the Google stock price example, the naïve model has no parameters, so $K = 0$

```r
# lag=h and fitdf=K
# Total lags used: 10
Box.test(res, lag=10, fitdf=0, type="Lj")
```

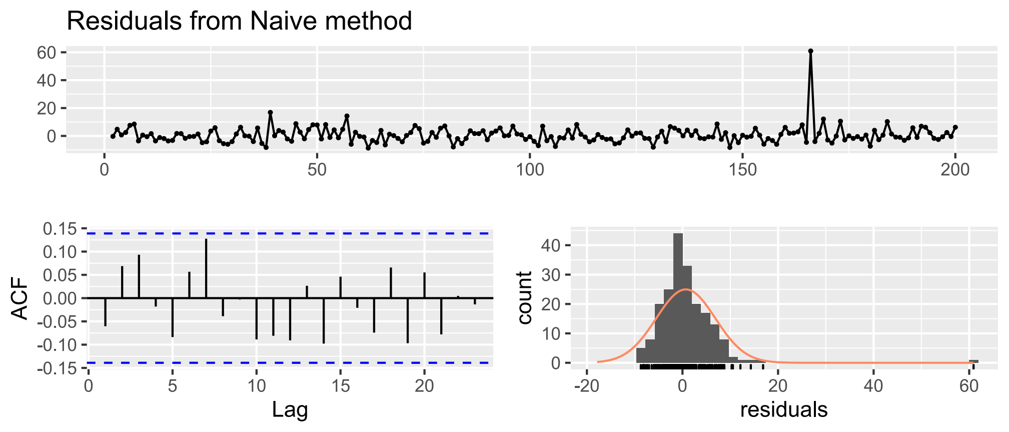
```
##
##      Box-Ljung test
##
## data:  res
## X-squared = 11.031, df = 10, p-value = 0.3551
```

- **The results are not statistically significant (i.e., the p-value>0.05). We fail to reject the null hypothesis H_0. Thus, we can conclude that the residuals are not distinguishable from a white noise series**.

Ljung-Box test (cont'd)

- Test if the residuals are white noise.
- `checkresiduals` function produces 1) a time plot of residuals, 2) a ACF of the residuals, 3) a histogram of the residuals (with an overlaid normal distribution for comparison), and 4) a Ljung-Box test.

```
checkresiduals(naive(goog200))
```



```
##
##      Ljung-Box test
##
## data:  Residuals from Naive method
## Q* = 11.031, df = 10, p-value = 0.3551
```

Chapter 4

ARIMA models

- The general **idea** of ARIMA models is to **capture autocorrelation**.
  - **Autocorrelation**: measure linear relationship between **lagged values** of a time series y.
  - For example, we measure the relationship between:
    - y_t and y_{t-1}
    - y_t and y_{t-2}
    - etc.
- **Major assumption**: stationarity.
- **Advantages**: strong underlying theory, flexible, etc.
- **Key concepts**: order, differencing.
  - Autoregressive Integrated Moving Average models (ARIMA ($\mathrm{p, d, q}$) models):
    - AR: $p =$ order of the autoregressive part.
    - I: $d =$ degree of first differencing involved.
    - MA: $q =$ order of the moving average part.

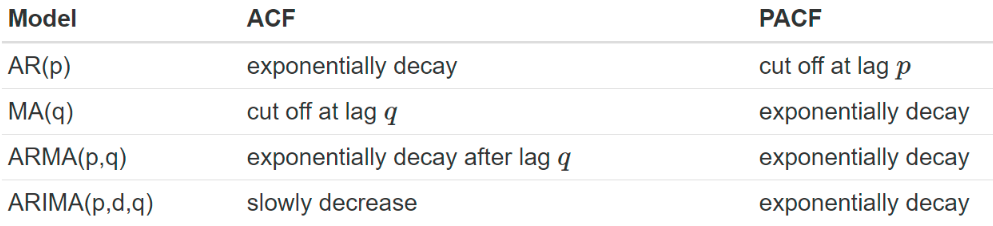
Stationarity

- First-order (Lag-1) differencing (used for removing trend)
    - The differenced series is the **change** between each observation in the original series:
    - $y'_t = y_t - y_{t-1}$. In R, `diff(data, lag=1)` or `diff(data)`.
- Seasonal (Lag-m) differencing (used for removing seasonality)
    - is the difference between an observation and the corresponding observation from the previous year.
    - $y'_t = y_t - y_{t-m}$, where m= number of seasons.
    - For monthly data m=12. In R, `diff(data, lag=12)`.
    - For quarterly data m=4. In R, `diff(data, lag=4)`.

Stationarity (cont'd)

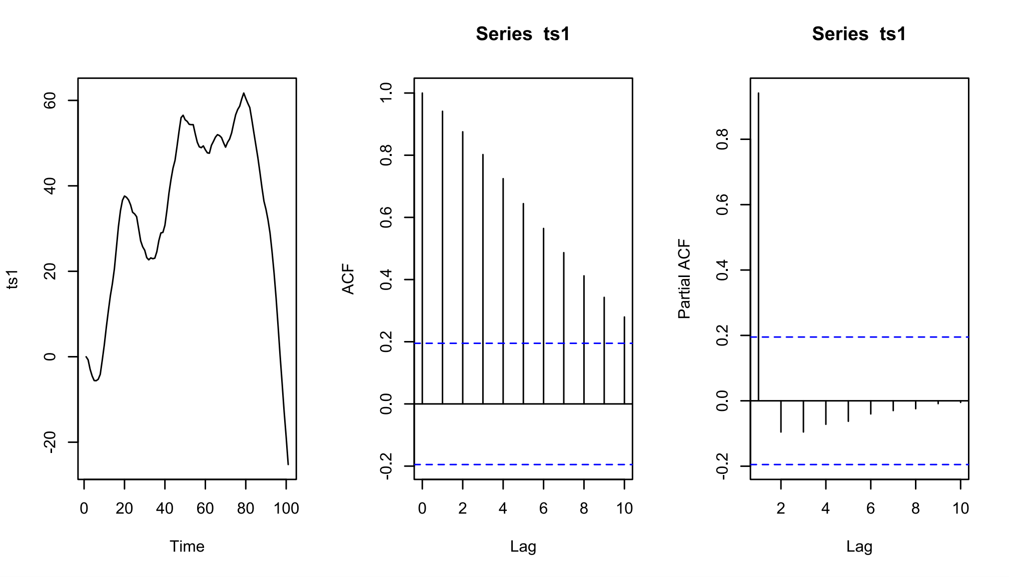
- Augmented Dickey-Fuller Test (ADF)
  - In general, the Augmented Dickey-Fuller test is defined as:
    - H_0 (null hypothesis): the data series is not stationary
    - H_a (alternative hypothesis): the data series is stationary.
- The `adf.test()` from the R package `tseries` will do a Augmented Dickey-Fuller test.

Identifying ARIMA (**p, d, q**) models



Model	ACF	PACF
AR(p)	exponentially decay	cut off at lag p
MA(q)	cut off at lag q	exponentially decay
ARMA(p,q)	exponentially decay after lag q	exponentially decay
ARIMA(p,d,q)	slowly decrease	exponentially decay

- ARIMA (1,1,1)
- Note: **ACF** starts at lag 0 and **PACF** starts at lag 1.



Seasonal ARIMA models

- (Non-seasonal) ARIMA models: non-seasonal data.

- ARIMA models are also capable of modelling a wide range of seasonal data by including additional seasonal terms.

- Seasonal ARIMA models:

ARIMA	$\underbrace{(p,d,q)}$	$\underbrace{(P,D,Q)_m}$
	↑	↑
	Non-seasonal part of the model	Seasonal part of the model

where $m =$ number of seasons/number of observations per year.

```
- For monthly data m=12.
- For quarterly data m=4.
```

Identifying Seasonal ARIMA models

- Identifying seasonal ARIMA (**p, d, q**) (**P, D, Q**) models:
    - The non-seasonal part of an AR or MA model will be seen in the lags of a PACF and ACF. For example, AR(p), a p-order of a AR part, cuts off at lag p of a PACF. MA(q), a q-order of a MA part, cuts off a lag q of a ACF.
    - The seasonal part of an AR or MA model will be seen in the seasonal lags of the PACF and ACF.

Chapter 5

Exponential smoothing

- **Exponential smoothing methods are**...
  - weighted averages of past observations, with the weights decaying exponentially as the observations get older.

Types of exponential smoothing

- What are the 3 main types of exponential smoothing?
  - **Simple exponential smoothing**: for series with no trend or seasonality.
  - **Holt's method**: with trend, no seasonality.
  - **Holt-Winter's method**: with trend & seasonality.

"All models are wrong, but some are useful."

--- George E. Box ---